DOCUMENT RESUME

ED 348 939 IR 015 58[illegible]

AUTHOR Becker, Henry Jay
TITLE Computer-Based Integrated Learning Systems in the Elementary and Middle Grades: A Critical Review and Synthesis of Evaluation Reports.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for Social Organization of Schools.
REPORT NO CSOS-48
PUB DATE Oct 90
NOTE 47p.
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Compensatory Education; *Computer Assisted Instruction; *Computer Managed Instruction; *Computer Networks; Educational Media; Elementary Education; Evaluation Methods; Individualized Instruction; Microcomputers; *Program Evaluation
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1; *Integrated Learning Systems

ABSTRACT

This paper reviews evaluation reports on the use of integrated learning systems (ILSs) in elementary and middle schools throughout the United States in light of the following questions: Do these systems work? Do students make more progress using an ILS than learning in "traditional" ways? Could the cost of software and equipment be better spent on smaller class size, higher teacher salaries, more print materials, inservice training in classroom management, or on training for using other innovative instructional programs? The following studies on the implementation of ILS are summarized: (1) the WICAT ILS in individual schools and Chapter I and compensatory education programs; (2) the use of Computer Curriculum Corporation (CCC) in Chapter I pull-out programs in public and parochial schools, including a 4-year study of CCC use in the Los Angeles (California) schools; (3) Jostens Learning Corporation's Basic Skills implementation; (4) a field experiment using computers in fifth through eighth grade mathematics in 31 schools; and (5) integrated systems piloted in New York City using CCC, WICAT, ESC, Prescription Learning, Wasatch, Ideal, Plato, CNS, and others. The analyses show that results differ as much based on different conditions of the study as on the different software packages used; students generally do somewhat better than expected using ILSs and sometimes the results are substantially superior; and evaluation data are not sufficient for use by a school district as a basis for acquisition and investment decisions. Tables showing comparisons of the various studies are included. (ALF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR

Report #48

ED348939

Computer-Based Integrated Learning Systems in the Elementary and Middle Grades:

A Critical Review and Synthesis of Evaluation Reports

Henry Jay Becker

Center for Social Organization of Schools

Johns Hopkins University

October, 1990

IR015580

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. H. Hollifield

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

Recent survey data shows that although most elementary school students use computers, that use has mainly been occasional and for purposes of lending variety and "enrichment" to the school day, rather than as a central component of teachers' instructional programs (Becker, 1990a). To a large extent that pattern of use is inevitable, given that most elementary computer-using teachers have only one or two computers available to use with their classes of 20 to 30 students. Furthermore, the loading of software into the vast majority (over 90%) of elementary school computers must be done individually, through separate diskettes or cassettes, and each diskette holds only a limited amount of programmed logic and educational content, particularly when much diskette space is consumed by motivational graphics and pictures.

In contrast to diskette-based software loaded into stand-alone microcomputers, some schools are now investing in "integrated," "individualized" computer software supplied by a single vendor and containing instruction and practice problems covering a multiple-year curriculum sequence. This software is housed on a central "server" computer linked in an electronic network to 15 to 30 student computers. Specific lessons are automatically loaded into each student's computer when that student "logs in" based on a continuous assessment of that student's previous accomplishments and current learning needs. Such systems are known as integrated learning systems.

Integrated learning systems (I.L.S.'s) actually pre-date the existence of microcomputers, but with the latter's rapid development, I.L.S.'s have increased in number, extensiveness, and sophistication. A recent report from the Educational Products Information Exchange (EPIE) has thoroughly assessed the characteristics and qualities of eleven such systems (EPIE Institute, 1990).

A strong rational argument can be made for why we should expect computer-based integrated learning systems to be effective--effective, that is, in helping us to improve student academic performance, particularly in basic mathematics and language skills. This argument has at least six components.

(1) **Computer-based** activities are motivational. The clearest empirical research that exists about instructional uses of computers is that students overwhelmingly enjoy most computer learning activities (Becker, 1986a).

(2) **Networked** computer-based activities eliminate some of the logistical problems that teachers face when using computers--for example, handling the plethora of diskettes and the time required to load and run programs at each student station.

(3) Networked computer-based systems with **centralized management** functions enable teachers, without spending very much effort or time, to target specific programs for specific students based on demonstrated needs or targeted curricular goals. A major use of this flexibility is for establishing a completely **individualized skill practice** through individualized placement and periodic re-placement, which provides a fast pace for self-teachable students, and more thorough review for slower students, giving them each a more appropriate level of instruction. Other possibilities for using management systems

are possible--for example, allocating assignments to small groups, providing records on student or group performance, and so on.

(4) Management systems with strong **diagnostic-prescriptive analysis systems** can provide assessments of specific skill deficiencies for specific students and bring those to the attention of teachers and other supervisors.

(5) Diagnostic-prescriptive analysis systems with strong **tutorial** capabilities can even provide necessary direct instruction to students with specific concept deficiencies.

(6) And networked systems supplied by **a single vendor** can provide a complete multi-level, multi-subject curriculum that presents a **consistent user-interface**, enabling students to concentrate on the substantive learning tasks rather than on figuring out how each specific program works.

It is important to note that not every package that presents itself as an integrated computer-based system provides all of these functions, and that there are other approaches to computer-based instruction that do use network-loaded software but are not packaged as integrated systems that also can provide many of these functions (Gevirtz and Kelman, 1990). Furthermore, the most appropriate choice of a so-called "integrated" system versus other networked-based approaches depends on one's assessment of the importance of other factors in the instructional delivery system--for example,

* the relative importance one places on having software that presents a consistent user interface versus the value of enabling teachers to choose from among alternative software products from different vendors;

* the value one places on teacher control and responsibility for successfully managing computer activities versus the value of relieving teachers of having to attend to those concerns;

* how well a particular I.L.S. vendor's materials addresses the particular curricular goals that one thinks are important;

* and the quality of lessons that is represented in the software package.

But one can endlessly debate the theoretical plusses and minuses of a single-vendor, centrally-managed, diagnostic-prescriptive, networked system for addressing curricular goals considered important by the school system. Still, the question remains -- do these systems work: do students make more progress using them than they would have had their time been spent learning the same material in more "traditional" ways? And is that progress made on curricular goals that are most important? Moreover, would the additional cost of equipment and software be better spent on smaller class sizes, on raising teacher salaries to attract more educated people into the teaching profession, on more print materials, on inservice training for handling classroom management problems or on training for using other innovative instructional programs?

Regardless of the theoretical argument for integrated learning systems, it is the empirical questions that are important, and it is those questions which this paper addresses with evidence.

The Universe of Evidence: Selective Dissemination

First, it is important to know where the evidence comes from. The studies summarized in this paper are a large subset of the following evaluation reports of integrated learning systems implementations:

* About 30 very consise, well-organized 2-page evaluation summaries of implementations of the WICAT I.L.S. system. The summaries were compiled by Wicat Systems, Inc. in a brochure called "Learning Improvement Reports" (Wicat, n.d.). These reports include a few district-wide evaluations, but are mainly reports of implementations in individual schools, including both schoolwide implementations and Chapter I and compensatory education programs.

* Five annotated statistical reports by the Computer Curriculum Corporation (C.C.C.) on the use of C.C.C. as a Chapter I pull-out program in three public and two parochial school districts (Suppes, Zanotti, and Smith, 1988a; 1988b; 1989a; 1989b; 1989c). In addition, I examined two well-done district reports of C.C.C. evaluations, one in a San Diego high school (Kitchen, 1986) and, the other, a National Diffusion Network-accepted evaluation in Lafayette Parish, Louisiana (Hotard and Cortez, n.d.).

* District analyses and a separate independent evaluation of the Calvert County, Maryland countywide C.C.C. implementation, including original tabulations of year-by-year results (Austin, 1988).

* The Educational Testing Service's report of the four-year long study of C.C.C. in four Los Angeles elementary schools in the late 1970's (Ragosta, Holland, and Jamison, 1982).

* Evaluations of C.C.C. programs in four districts (Rochester, Atlanta, Newark, and Portland), which I had examined in an earlier literature search (Shattuck, 1986; Atlanta Public Schools, 1984; Gourgey, et al, 1984; Miller, 1984; examined in Becker, 1986b).

* Reports of seven Jostens Learning Corporation's Basic Skills implementations, three as summaries produced by the company several years ago, one in the form of a few statistical tables provided by a company marketing representative, two L.E.A. reports (Zollman, Oldham, and Wyrick, 1989; Sacramento Unified School District, 1989), and a doctoral dissertation using an expermental design in a Fairfax, Virginia elementary school (Alifrangis, 1989).

* Data from my own field experiment on using computers in fifth- through eighth-grade mathematics conducted in 31 schools, including one school that used the Jostens program (Becker, 1990b).

* An evaluation report on a number of integrated systems piloted in New York City Schools, including sites with C.C.C., Wicat, E.S.C., Prescription Learning, Wasatch, Ideal, Plato, C.N.S., and several others (New York City Public Schools, 1989).

Although these studies represent a diverse array of district-disseminated and vendor-disseminated evaluations and independent research projects, they cannot really be seen as an unbiased sample of current implementations of integrated systems. Apart from the fact that schools and districts that go to the effort to conduct an evaluation are likely to be somewhat better organized and to conduct more careful implementations than sites that do not evaluate, the studies that are disseminated are not likely to be representative of the data collection efforts that go on. The effort that goes into publicizing a particularly unfavorable or even negligibly effective study must be less, on the average, than the effort taken to disseminate the results of a successful experiment. This must be particularly so when large financial investments have been made, as where a school district makes a central command decision to purchase an I.L.S.

Thus, the first barrier against using research data to understand the effects of computer-based instructional programs is the substantial bias in what information makes its way into the public debate. The impact of selective dissemination on the conclusions that we draw can be enormous. Consider these results from one of my own studies--a nationwide field experiment on computer use in grade 5 to 8 mathematics, a study where I.L.S.'s played a relatively minor role. In that study, an experimental design was in place in 29 schools such that one or more teachers taught one math class using computers and another class of similar-ability students using traditional methods. On all five outcomes that were assessed, the average effect size was negligably different from zero. For example, on the Stanford Achievement Test's math computation subtest, the average effect size was a disappointingly small +.01--meaning that the progress during the year in the computer classes averaged only 1/100 of a standard deviation more than did progress in the traditionally taught classes.

But suppose that the only schools that publicized their results were those that got positive results--ANY positive results, even by the whisker-thinnest margin above zero. Then, we would conclude that the computer treatment had a fairly positive effect. The average effect size on the math computation subtest for that group of schools was +.28, meaning that compared to very comparable traditionally-taught classes, the computer-using classes gained about 6 NCE points or, roughly, at the middle range of performance where these schools are, 11 percentile points. That difference--between no effect at all and a reasonably substantial but not earthshaking one--could conceivably be the size of the bias that we observe when we look only at studies disseminated by vendors. To draw accurate conclusions, we clearly need advance commitments of districts to publicize the results of their studies, regardless of the size or direction of those results.

Selective dissemination is only one barrier to a fair interpretation of existing data on I.L.S.'s. I encountered at least three other major problems as I examined the data. The three problems that I encountered were (1) poor evaluation designs compounded by inadequate data collection; (2) poor data analysis and presentation of results; and (3) inadequate description of the implementation (how the program operated) and the conditions and environment where the program was used.

Design and Data Collection.

To judge the difference that an I.L.S. has made on student achievement, most administrators understand the need for making a fair comparison between test score gains made by I.L.S.-using students and gains made by other students. Of course, not just any other students will do. The question that we are trying to answer is what would those same students who used the I.L.S. have accomplished had they not used it. Experimental designs, where some students from a specific population--for example, Chapter I students in a district--are randomly assigned to use an I.L.S. while others are assigned to other structures for remediation, provide the best way to create a comparison group that will most resemble the students using the I.L.S. Unfortunately, few program implementations are planned with evaluation strongly in mind, and evaluators have to scurry about trying to identify an appropriate group of students whose achievement data can be compared with that of students using the I.L.S.

In reviewing the studies on I.L.S.'s, I found four sources of comparison data that were usually used:

* data on the same students (or students in the same class cohort) from earlier in their school lives--longitudinal studies;

* data on students from the same schools, but who passed through the tested grade level in previous years--cohort comparisons;

* data from other schools nearby--comparison school studies;

* and data on the entire population of students who took the same achievement tests--the norming population.

Unfortunately, there are problems with all of those methods, and they profoundly affect our ability to make strong conclusions about the existing data. For example:

* Comparisons with scores or gains made by another school down the road, or with a previous year's cohort are suspect when there is evidence that something is not typical or comparable about that comparison population. There are times when a comparison school or cohort sometimes shows such weak results that any program would look good in comparison. It is often more appropriate to use a larger comparison population (e.g., several years of previous cohorts or all the students in a district), but even a comparison with the entire district sometimes produces questionable conclusions, as I will illustrate.

* When a group of students is followed over time, usually evaluators look at group means: how the 5th grade class did as 3rd graders two years earlier. But that 3rd grade class was different. In the meantime, many of the lower-performing students have been held back, to be replaced in the class by older students held back a year earlier who are now doing the grade level's work for a 2nd time. Those repeaters may not be the top students in the class, but they will clearly do better than the failed students that they replaced.

* But the most frustrating problems come from inadequately normed standardized tests. This shows up most sharply as Fall-to-Spring gains, and the problem is created by not having a comparison group tested during a similar interval. Students take a test in September. They score at the 50th percentile. They use an I.L.S. throughout the year; take another test in May; score at the 60th percentile. Is this evidence of an "effect"? Not necessarily. For reasons that are unclear, students typically make 10 percentile point gains from fall to spring, even when season-specific norms are used in the calculations. Again, consider data from my own study. (See Figure 1.) The solid line shows the distribution of mean fall-to-spring percentile gains of students among 27 computer-using math classes, from the class that made the least gain (actually a substantial decline) to the class that made the greatest gain. The median class-average gain is just under 10 percentage points. The dashed line in the same figure shows the distribution of mean percentile gains for 27 traditionally taught classes, classes that did not use computers, and where each class was matched with a computer-using class on prior student achievement, and in most cases was taught by the same teacher. The median for the traditionally taught classes is exactly the same--just under a 10 percentile point gain. It could be that these are wonderful teachers who taught well regardless of having computer software or not; or that they learned some terrific new teaching methods that they imported from their computer-using class to their traditionally-instructed class. But it is more likely that the test norming for fall and spring testing points is just not equivalent. This conclusion is supported by other analyses of test score data from Chapter I programs, where it is typical to find N.C.E. gains of 8 points, which, for a relatively low-scoring population is also about a 10 percentile point gain.

Figure 1 about here

And fall-to-spring gains are not the only ones hampered by a lack of appropriate norming. On Spring-to-Spring comparisons, nationally, studies have found students and grade cohorts averaging gains between 1 and 2.5 N.C.E. points per year, with lower-performing students at the higher end of this range (Gabriel, et al, 1985; Anderson, 1984). Because most district and school officials heavily value high test score performance, administrators and teachers are increasingly providing instruction in test-taking skills, orienting the taught curriculum and materials increasingly oriented towards identical skills and procedures as those used on the standardized tests, and may be engaging in questionable administrative practices as well (e.g., allowing students a longer test period, excluding special education students, or discouraging students with attendance or achievement problems from attending school on testing days). So without a comparison group, there are a number of grounds to question or even reject studies that are based on Spring-to-Spring gains of a specific cohort or changes in successive cohorts for the same grade.

There are other design and data collection issues besides identifying appropriate comparison data, but I only want to mention one: the appropriateness of the achievement test to measure program-related consequences.

Nearly all of the evaluations reviewed used nationally normed standardized achievement tests such as the Iowas, Californias, CTBS, Stanfords, and so on. A few used statewide competency tests. And several used tests internal to the I.L.S., but none used those same tests with a comparable student population from the district that did not receive the I.L.S. treatment. In my own study, included in the analysis because two sites did use I.L.S.'s, student achievement was measured in a second way in addition to standardized tests. For each pair of classes (computer-using and traditionally-instructed), I developed a test built from computer program exercises and from other instructional experiences of each class, but overlapping as little as possible with the standardized test. I attempted to make the test "fair" so that both classes compared had roughly the same exposure to the test items as a group. These curriculum-specific tests, although not as sophisticated psychometrically as standardized tests, provide the benefit of more being more closely matched to the actual instruction and practice experienced by students in each specific pair of classes.

Data Analysis and Presentation.

Problems with how evaluators analyze their data and how they present it also produce headaches for a reviewer of the evidence. Each study uses different statistics--grade-equivalents, percentiles, raw scores, scale scores, normal curve equivalents. Many reports are weighted down with peripherally relevant data--like statistical significance tests on large samples or tests of homogeneity of regression coefficients--and fail to report an answer to the central question--how much of a difference in performance growth was there between students experiencing an I.L.S. and a comparable comparison sample of students. And some reports have unexplained anomolies--test results on substantially fewer students than were apparently present in the school or classes using the program, or containing results for only selected grade levels or selected test subjects.

To make some sense out of the diversity of statistics shown in these reports, this analysis converts the different types of information about effects presented in each report into a common measure of effectiveness. That measure, now widely used in educational research, is called an effect size (E.S.). It is computed, simply, as the difference in performance between (in our case) the I.L.S.-using students and a comparison group, expressed in terms of the proportion of one standard deviation of individual performance scores in the comparison group. Because I.L.S. and comparison students might have had different initial achievement, "growth" scores or other ways of adjusting "post-test" achievement measures for differential "pre-test" scores, are employed in many analyses.

What is a large E.S.? Most people would consider it to be a major accomplishment to take a fairly homogeneous low-scoring school whose students averaged at the 25th percentile in one year to a 35th percentile average score the next year, and to the 45th percentile in 3 years, with no change in the social makeup of the school. Assuming NO upward drift of the national test-taking population, (and assuming a comparison group post-test standard deviation of 15 N.C.E.'s, 71% of the standard deviation of a nationally heterogeneous population), that would be an E.S. of +.4 for one year and about +.75 for the 3 year period. But because there has been a national upward drift of standardized test scores--recently, all states reported means above the 50th percentile (Cannell, 1988)--it would be plausible to assume that in the absence of a special

program our hypothetical school's scores would have risen anyway. Based on an average of three studies--Gabriel, et al, 1985; Anderson, 1984; and the Maryland state data used in this paper--we might expect to see over three years a gain on the average of 4.5 N.C.E. points. That would put the comparison population at the 33rd percentile after three years. If we calculate an effect size based on this "moving" comparison point, the one-year effect size for this hypothetical school instructional program would be +.3; over three years, it would be +.47. Given an appropriate comparison group, those seem like plausible points to mark a "successful" large-scale instructional intervention.

Description of Program Implementation and Environment.

A fourth general problem in computer-based instruction evaluation reports is that they typically devote litte attention to describing how the program was implemented and under what conditions. Such descriptions are essential for drawing useful interpretations of the statistical summary of any particular implemention. We would want to know, for example, what training, education, and expertise was evidenced by the lab manager; how successfully was the schedule adhered to and to what extent other events (assemblies, other duties of the lab manager, downtime, teachers just having other plans) interfered with regular use of the lab; how often were the teachers present and what did they do in the lab (or what instruction did they provide away from the lab during lab time); what was the culture in the lab--noisy, over-controlled, etc.; how much, if at all, were teachers' lessons and reviews and individual tutoring affected by what they saw in the lab or by the informational reports they were given; and so on. Data about how instructional time was allocated across different subjects is also important to have.

Even more overlooked than descriptions of the implemented program, however, are descriptions of the environment which is being compared to the I.L.S. implementation--the "other" school, other cohort, or other class. For example, one of the strongest potentially confounding factors in analyzing the effectiveness of an I.L.S. program is knowing whether the students using the program spent more school time during the year on the subjects for which they were tested than did students in comparison schools, years, or classes. If the computer program was implemented as an add-on, it becomes difficult to know whether it was the program that made the difference or merely the additional time-on-task.

Heterogeneity of Results

The evaluations that were reviewed vary from studies of fewer than 10 students to studies of several thousand. Some were conducted in middle-class settings, others in the poorest of inner-cities. Some involved the youngest elementary students and others were of high school students or adults. Although all focused on basic skills learning and all were measured on tests that tested those skills, variations in content and test characteristics were still present. And different methodologies, design, data collection, and data analysis problems, and research styles governed their production.

There is some value in finding means or average effects among such studies. For we cannot carry around in our heads hundreds of conclusions that apply only to specific instances

and wait until we find the one instance that perfectly matches our specific need. Averages also tend to focus the conclusion on what is common about the implementations, leaving other causes that are due to happenstance--a new vibrant principal, a disorganized setting, hardware malfunctions, and so on--to cancel one another out, disappearing among the large selection of data. On the other hand, it is also easy to err in the other direction--to combine sets of studies of dissimilar environments important in their own right and for which different conclusions might be drawn about the effectiveness of I.L.S. programs.

In what follows, I have chosen to focus only on studies of elementary and middle-grade populations; and only those that serve either all or most of an entire grade-level in a school or those that serve as Chapter I pull-out programs for a substantial number of students. I have omitted studies of high-school-age remediation and dropout prevention programs, studies of adults, and programs dealing with fewer than 50 students across all grades. Also, with some exceptions, I will not comment on studies for which there was an absence of any point of comparison whereby an effect size could be at least roughly estimated, providing a standard metric to judge the achievement of the I.L.S.-using students with students using other instructional programs or no program in particular. As a result of applying these criteria, we omit further examination of 14 of the studies in the WICAT Learning Reports, one of the five vendor-supplied CCC studies, two of the three older vendor-supplied reports from Jostens, and one independent study of the Jostens program.

Table of Effectiveness of Elementary and Middle Grade I.L.S. Programs

Table 1, beginning on the next page, is a compact summary of each of the I.L.S. research reports re-analyzed for this paper and a corresponding summary of the re-analysis of each report. The reports are sorted by vendor and, within vendor, similar studies are grouped together. The table begins with 16 evaluations of the WICAT system. That is followed by 11 evaluations of the C.C.C. program (4 vendor evaluations and 7 done by the districts involved or independent researchers). Following these are 4 evaluations of the Jostens (E.S.C.) integrated system, and a multi-vendor evaluation done in New York City.

Table 1 (all 10 pages of it!) about here

Each entry contains six columns of data. The first column indicates the location of the implementation (school, city, state); the specific I.L.S. system used; who conducted the evaluation and who produced the report that describes the results (vendor, school system, independent researcher); and a capsule categorization of the type of research design and comparison group used in the original evaluation. The second column indicates roughly how many student-users of the I.L.S. system were included in the evaluation (i.e., computer-using students for whom achievement test data were used to assess effectiveness); the students' grade levels; socio-demographic characteristics of the school or student population; and how extensive an I.L.S. experience the students had (number of hours during which they used the system for the subject(s) on which they were tested or number of clock-hours recorded by the system during which they were logged in).

The remaining columns present several statistical indicators of how effective the I.L.S. implementation was in each study. Column three summarizes how the original evaluation described the effectiveness--how achievement was measured, what comparison among students was made, and what was found. Column four takes the data as provided by the original evaluator and translates the evaluator's comparison into an effect size statistic, showing, where necessary, what assumptions were made (principally concerning the standard deviation of the comparison group). Results are bracketed between high and low estimates or shown as estimated values (~) where assumptions had to be made.

Columns five and six present this researcher's analysis of the original research design, first, a summary of problems caused by drawing conclusions directly from the evaluator's statistics, and then, where appropriate, an alternative analysis of effects and its summary in terms of effect sizes.

The next section of this paper reviews the studies contained in Table 1.

<u>WICAT: The Chicago studies.</u>

WICAT included in its "Learning Improvement Reports" evaluation data for four separately studied Chicago elementary schools (WICAT Systems, n.d.). All were Chapter I

schools serving mainly minority residents of public housing. The reports indicate that in all four schools there was a substantial improvement in year-to-year standardized test scores between prior year-to-year gains and gains during an I.L.S.-using school year. Based on vendor-supplied summaries, effect sizes (E.S.) averaged about +.3 to +.4 for a single year. Average exposure to I.L.S. materials in these schools ranged from an estimated 35 class hours to 100 class hours although the vendor's data show similar effect sizes regardless of time of computer use.

In two of the schools (Beale and DuBois--see Table 1), score gains were reported for a very small number of students (in Beale, for 50 students in the two-lowest-performing Chapter I classes in grades 5-6; in DuBois, 56 students across grades 1-8), raising the issue of why results for those specific students were reported but not results for larger numbers of students. (WICAT apparently reported whatever data were provided to it by the school, but achievement data on other categories of students may have been collected but simply not reported by their respective school.)

In the other two schools (Colman and McCorkle), the score gains were reported for hundreds of students, but the positive effect sizes in those schools are based not on substantial year-to-year gains during the I.L.S.-using year, but on extremely low gains during the prior (comparison) years. For example, in Beale and DuBois, the two other Chicago schools studied, students gained an average of about 7/10 to 3/4 of a grade equivalent per year prior to their use of computers, a level typical for inner-city schools. But in the pre-computer years at Colman and McCorkle, students reportedly gained only about 4/10 of a grade equivalent per year, a level lower than likely to exist at any other regular public school in the country with normal test administration and scoring conditions. So it is plausible to think that test scores were systematically depressed at Colman and McCorkle prior to the establishment of the I.L.S. program and that those "pre-" scores do not represent "normal" scores for their student populations. Perhaps the schools had been very weakly managed with little test-taking preparation or control of student behavior on test days; perhaps an unusual number of errors were made in handling student data. The implementation of an I.L.S. program at these two schools may have been part of a general staff and administrative housecleaning. We have no basis to know that this was so. But we do know that inner-city schools, particularly those encompassing grades 7 and 8, are very unlikely to average lower than 6/10 grade equivalent gain per year, let alone the 4/10 that was reported. If, instead of the reported comparison point, we use a more typical inner-city school mean grade equivalent gain of 3/4 year per calendar year, effect sizes for Colman and McCorkle become negligible, ranging from about -.1 to +.1.

WICAT: The Texas studies:

The Learning Improvement Reports included five reports from Texas districts, 3 of them evaluations of a single school and 2 of them evaluations of district-wide implementations involving thousands of schools. The outcome measure for these evaluations was a statewide competency test given each February. Evaluations were made in terms of changes over time in the percent of students passing each test (reading, writing, and mathematics). But whereas the Chicago studies used grade-equivalent gain scores of prior cohorts at the same school as a comparison point, the Texas evaluations compared changes from cohort to cohort in the percent passing the test (from pre-computer to post-computer) in the I.L.S. school or district with changes

during the same period in another student population. Three comparisons were made with changes in statewide scores; one with changes in the district in which the school was located; and one with a single neighboring school. WICAT reported results which, translated to effect size terms, indicate minor effectiveness in two suburban, middle-class settings, modest positive effects in two Chapter I schools, and large positive effects in a 90% Hispanic school district.

Questions can be raised, however, about the appropriateness of each of these calculated effect sizes. In both cases where a comparison group other than the statewide average was used, the comparison group (change in percent passing) performed rather poorly in comparison with the state average. Statewide changes in percent passing the test would seem to be a better point of comparison than changes in a district or in a neighboring school. The latter are subject to small-scale year-to-year cohort variations or variations in local circumstances. The best comparison point for the Chapter I schools and districts would seem to be statewide data on similar schools, but unfortunately that data is not present for any of the evaluations.

Of the three Texas studies that used statewide averages as a point of comparison, two reported a very peculiar pattern of effects. Recall that the testing period is in February, so that in the first year of implementing the program, only four to five months elapsed prior to the "post-test" (including start-up time, which often has built-in inefficiencies while problems are ironed out). One would expect that effects might be quite modest for that first period, but be more substantial during the following 12 months (i.e., between the latter half of the first year and the middle of the second year). In fact, in both cases, most or all of the effectiveness occurred during the first few months of the intervention. The district-wide program, for example, had an effect size of +.47 for the first half-year, but an effect size of -.05 for the full year following. It is possible that the district had particularly depressed scores in the year prior to the implementation of the I.L.S., perhaps moving the district to take major rehabilitative steps including, but not limited to, I.L.S. implementation.

WICAT: Other studies.

The seven remaining WICAT studies are more heterogeneous than the above groups of studies, although all seven provide data on between 80 and 300 computer-using students, a relatively small number. Translating results reported by WICAT to the effect-size metric yields portrait clearly above +.3, but most evaluations suffer from at least the problem of assuming that national norm populations are constant from year to year, when in fact, there has been a substantial rise in standardized test scores across the nation and for other large comparison groups. At least two of the studies' positive conclusions must be disregarded for that reason, and the effect sizes of at least two others are reduced by taking national norm upward drift into account. Still, a majority of these implementations seem to have been effective, although without knowing more about each setting, one cannot be sure that results were not due to other factors such as changes in educational services coincident with the implementation of an I.L.S. system. No WICAT study employed a randomized design, comparing a population of WICAT users with a known-to-be-equivalent population of students who did not use WICAT. Of the seven studies, only the Memphis study in Catholic, middle-income schools provided some evidence on the comparability of comparison populations and provided data on changes in achievement on the same students in both treatment populations. But even there, one aches to know more about the

comparison schools, and whether their relatively poor results might be at least partly responsible for the positive effect size attributed to the I.L.S. system.

C.C.C.: Vendor-supplied studies.

By far the largest reported effects among those reviewed were from these set of four evaluations supplied by C.C.C. (Suppes, et al, 1988a; 1988b; 1989b; 1989c). Each was a year-long study, involving between one-half and one hour of use per week, distributed among math and reading activities. Two were modest in size (under 200 students); two others were larger (about 600 students). Effect sizes computed from vendor data were astronomically large, ranging from +.6 to +1.6. With effects of that magnitude, incorporating a correction for national upward drift of test scores made little difference--even using a correction factor of 2 NCE per year, which appears to be typical of Chapter I populations such as those in all of these C.C.C. studies (Gabriel, et al, 1985; Anderson, 1984).

However, before concluding that the C.C.C. program had such large effects, we must draw attention to two analytic problems peculiar to this set of studies. Vendor evaluators used the unusual procedure of eliminating cases from data analysis that showed sharp declines, under the assumption that singular conditions rendered certain individuals' posttests to be atypical portraits of their abilities. However, students with correspondingly large gains stemming from singularly low pre-test scores were not excluded. This asymmetry, defended on the grounds that the treatment might have caused the large gain, but could not have caused the sharp decline, ignores alternative explanations for large gains that are the same as those for sharp declines--peculiar circumstances on one test-taking day for an occasional individual. Incorporating a correction for this asymettric selection of cases reduces the measured effect size substantially only for one of the four studies. But even that study yielded a (corrected) E.S. of +.35.

Two of the four studies are also questionnable in terms of a high attrition rate: test scores are provided for only 60% of the students using the program. Presumably the others all left the school system during the year, and posttest scores were not available for them. But it would be instructive to know whether the dropouts differed systematically from those who remained.

Unfortunately, the substantial effects reported in C.C.C.-provided evaluations are not replicated in other evaluations of the C.C.C. program, including two randomized experiments and some district-wide evaluations involving larger numbers of students. We turn to those next.

C.C.C. in Calvert County, Maryland

In this county-wide district, substantial increases in achievement test scores during the 1980's have led to a great deal of vendor- and school district publicity about the effectiveness of C.C.C. District administrators proudly display colorful graphs showing steep slopes of mean test scores. Test scores are reported (accurately) to have risen even in comparison to other countywide districts in the state. "Calvert County's achievement test ranking in the state rose from fourteenth in 1979-80 to third in 1986-87" (Austin, 1987). The district summarizes:

> "Schools in Calvert County are demonstrating what a total commitment to computer-assisted instruction can do for children. Student performance in the basic skill areas of reading, language arts, and mathematics is significantly improved with the use of this instructional tool. Better than 83 percent of the students are scoring about the 50th percentile, and the first three stanines in reading and the first three stanines in mathematics have been virtually eliminated." (Calvert County Public Schools, n.d.)

However, neither district reports nor independently conducted research draws appropriate conclusions regarding the effectiveness of the I.L.S. program in Calvert County.

The program was instituted in elementary schools county-wide in November of 1983. All students received regular math, reading, and language arts instruction. It was extended to the middle grade schools about January, 1985. However, district reports typically compare their most recent scores with scores from the Fall, 1980 testing, which was *three years prior to the onset of I.L.S. use*. Examining a more complete set of state and district test score means (grade equivalents) for every fall since 1980 is particularly instructive.

Both district and state mean grade equivalent scores were rising between 1980 and 1986 (Table 2). And they were rising in Calvert County faster than in the state as a whole. But the differential rise in Calvert County (over the state gain) was even larger prior to the implementation of I.L.S. than it was afterwards. And since 1986, the district has fared worse than the state average. At all three grade levels tested (3, 5, and 8) district scores are worse as of 1988 than they were at the onset of the I.L.S. program, after statewide gains in mean grade equivalent test scores are taken into account.

Table 2 about here

Another way to look at this data is to see whether the grade cohorts that had several years of C.C.C. made greater test score gains than the earlier cohorts that did not use the program, again using gains statewide of the same cohort as a comparison point. That analysis shows that between the 3rd and 5th grade, four cohorts that had two years of C.C.C. did slightly better than two earlier cohorts that experienced 3rd and 4th grades before C.C.C. (E.S. ~ +.12). However, between the 5th and 8th grades, two cohorts that had 2.5 to 3 years of C.C.C. showed less improvement in scores than three cohorts that had at most one-half year of C.C.C. (E.S. ~ -.25).

Even the slight gain of the younger students might be attributable to the substantial demographic changes occuring in the district during the 1980's, with younger, more educated suburban families moving into a largely older, rural district. More recent cohorts are thus composed of higher scoring children who are also likely to be making greater progress from year to year.

In order to take such demographic changes into account, Austin and Howie (1990) did a longitudinal analysis of individual students of two grade cohorts, those who were in grades 2 and 3 in 1982. The students' test scores were followed for 5 years, two years prior to C.C.C.

starting and three years afterwards. The researchers measured the effects of the I.L.S. program by comparing "pre" and "post" Normal Curve Equivalent (N.C.E.) scores using the average of students' two "pre" years' scores as the baseline and using the average of their three "post" years' scores as the outcome measure. They found that the mean gain of students over this 2.5 year period was about 5.3 N.C.E. units. (It was over 7 units for the younger cohort and under 4 units for the older one.) However, that analysis is flawed in two respects.

First, there was a distinct systematic upward drift of student N.C.E. scores between the two years prior to starting C.C.C. For example, data on one of the two student populations (given in percentiles rather than N.C.E. units, in Austin, 1988) shows the following median scores from pre-year-1 to post-year-3 on the total achievement test battery: 52, 66, 65, 68, 67. For this population, essentially all of the improvement (relative to the normed population) was between the first and second pretest years. (For the other student population, the improvement between the first and second pretest years was roughly equal to the total improvement between the second pretest year and the average of the three posttest years.) The effect of averaging two years of pretest scores, rather than using only the single year just prior to program implementation, was to incorporate a systematic downward bias of measured pretest scores.

In addition, the researchers did not take into account gains in the average N.C.E. score made by the same grade cohort statewide between the immediately pre-C.C.C. year and the "average" post-C.C.C. year, gains which can be estimated (from published grade equivalent means for 1983 and 1985) to be roughly 3 N.C.E. units for the 3rd grade cohort of 1983 over the two years beginning in 1983.

If one makes both adjustments--i.e., uses only 1983 data as pretests and corrects for a 3 N.C.E. statewide improvement between 1983 and 1985--the corrected effect size becomes negative, roughly E.S. = -.1 (The actual estimate calculated, given a standard deviation of 15 NCE, is -.13.)

The researchers also noted that the 94 Chapter I students in their populations made substantial NCE gains between the averaged "pre" and averaged "post" years. However, when the same adjustments are made for that population, the effect size for the Chapter I students also becomes negative.

Thus, a reexamination of both district-level cohort test scores and longitudinal data on individual students suggests an alternative conclusion to that which has been widely publicized--that test scores in Calvert County between 1983 and 1988 rose no faster than they would have in the absence of an I.L.S. program.

C.C.C.: Two Randomized Experiments

All studies discussed to this point have involved essentially the manufacturing of artificial comparison groups that might stand as standards for how students in an I.L.S.-using school would have fared had they not used the computer-based instructional program. Analysis of that kind of data, as I have shown, is subject to misleading conclusions based largely on the use of comparison groups of questionnable validity. The solution to the critical problem of appropriate

comparison groups is to engage a school or district in a randomized experimental design, with students assigned by chance to I.L.S.-using and traditionally-instructed (non-using) groups. Two studies of C.C.C. I.L.S. products did just that. One C.C.C.-using experimental design was a small study at two Chapter I schools in Louisiana in the mid-1980's. The other one was a larger study, conducted over three-and-a-half year period in four Los Angeles schools more than 10 years ago. The L.A. study used between-cohort and comparison-school designs and an experimental design all over the same experimental population. The form that the experimental design took was to have equivalent sub-samples of students use computers differentially for mathematics, for reading and language arts, or for both subjects. For any one subject area, students either had computer work or they did not.

The Los Angeles study, conducted by the Educational Testing Service, was fully reported in Ragosta, Holland, and Jamison (1982). A summary of their findings can be found in Ragosta (1983). Except for this author's own study (see below), the E.T.S. study was the only study of a single I.L.S. to use the Effect Size as its measure of differential achievement outcome. Table 1 contains some of their reported results, averaged over multiple grade cohorts and subtest outcomes. Generally, E.T.S. reported substantial effect sizes on standardized achievement tests for mathematics computation, accumulating to +.7 over three years. Effects sizes for math concepts and applications were close to zero, making the total mathematics scores somewhat lower, but still substantially positive. Modest but clearly positive effect sizes were also reported for reading and language, although positive results in these areas were more consistent for a single-year intervention than for a three-year duration. E.T.S. also reported huge positive effect sizes for a battery of curriculum-specific tests taken directly from their computer-based practice activities. However, that battery was not "balanced" by an attention to material experienced by the "control" group classes but not by the computer-using group. Moreover, because instructional time in the measured subjects was not equal between experimental and control groups, all test results are somewhat biased in favor of the experimental groups.

Most importantly, though, all E.T.S.-reported effect sizes must be adjusted downwards because they used a non-standard and upwards-biasing standard deviation measure in all effect size calculations. Their effect sizes are based on a standard deviation of residual posttest scores after controlling for pretest scores (and other background factors). Because pre- and posttests are highly correlated, the distribution of residual posttest scores is more compact than the actual raw posttest scores. The problem is worse for reading and language arts than for math. Re-calculating effect sizes based on this revised (but more standard) definition of standard deviation, I found that individual mathematics effect sizes were diminished by 25% and reading/language results by 40%. The final column in Table 1 reports the mean revised effect sizes for the various grade-level/cohort data points that the original researchers used. These show still substantial effects for mathematics computation, and still positive, but somewhat reduced effects for total math, reading, and language arts, for both cross-cohort and experimental comparisons. Specifically, the re-calculated effect sizes are +.26 for math computation over one year and +.59 over three years; +.17 and +.27 for total math (1 and 3 years respectively); +.12 and +.06 for reading; and +.11 and +.14 for language arts.

The relevance of the L.A. study for understanding of the current C.C.C. product is only modest, though, because the evaluated implementation was of a much earlier version of the software than is now being marketed to schools. (Actually, none of the research on C.C.C.

incorporates color graphics and the updated curriculum content that is part of the current product.)

The Lafayette Parish experiment, done roughly the same time as the Los Angeles study, found an effect size of +.2 for one year of mathematics implementation, roughly the same result as was observed in L.A. (once the standard deviation measure is adjusted to the more standard definition) (Hotard and Cortez, n.d.).

C.C.C.: Other Studies

Table 1 contains summaries of three other district-produced studies of C.C.C., but all date from the mid-1980's (Atlanta Public Schools, 1984; Gourgey, et al, 1984; Miller, 1984; Shattuck, 1986). None of these were experiments, and they provided variously good to poor data on the match between comparison students and computer users. Results were varied as well, and none are as useful for understanding likely effects of that C.C.C. program as the L.A. and Louisiana studies.

Studies of the Jostens Learning Corporation Basic Learning System

The Jostens Learning Corporation's Basic Learning System is represented in Table 1 by four studies. Data from two small studies were supplied by Jostens, and two independent researchers (one being this author) conducted two experimental studies involving the Jostens product. Neither of the vendor-supplied studies were based on very good designs, but if one had to draw conclusions from them, one would say the results were mixed--one implementation was likely to have been effective; the other was not (Sacramento Unified School District, 1989; Zollman, et al, 1989). Jostens has recently awarded a grant to an independent university-affiliated research organization to conduct large-scale systematic evaluations of its Basic Learning System in several different districts. In addition, Baylor University is leading a consortium of Texas districts in year-long evaluations of the Jostens system during 1990-91, and a number of these involve randomized designs. And this author has recently completed fieldwork on a small experiment involving the Jostens system at one elementary school.

The two completed researcher-done studies of Jostens came to two different conclusions. The Fairfax, Virginia study used an experimental design much like the Los Angeles C.C.C. study: students in grade levels four through six were individually and randomly assigned either to mathematics or to reading for the full school year. Each group thus served as a control for the other. Evaluation was in terms of changes in raw (or standard) CTBS scores from fall to spring. Very small differences favoring the I.L.S. were observed (Alifrangis, 1989).

In my own study, in which the Jostens experiment was only one of about 30 different experiments on the use of computer software in middle-grade math instruction, the results were quite different. This experiment occurred in a suburban area near Dallas, Texas over two years with two teachers, a fifth-grade and a sixth-grade teacher. Each teacher taught two math classes, and students from the bottom two-thirds of each grade cohort were randomly assigned to the "computer" or the "traditional" math class taught by the same teacher. Posttests included both

the Stanford Achievement Test (computation and applications) and a curriculum-specific test for each pair of classes comprised of items from assignments of both the "computer" and "traditional" class of any given pair. Fall pretests served as controls in a regression analysis of effect sizes.

Over the four computer and four traditional classes, one-year effect sizes were extremely positive. (Students were re-randomized for the second year, so two-year effects could not be measured. There was also substantial within-year and between-year attrition, dampening the useful sample size.) The mean effect size for total mathematics was +.45 on the standardized test and +.49 on the curriculum-specific tests. A separate "estimation/mental math" test, researcher-developed, had an even larger effect size (Becker, 1990b).

Although this was a small study, it was one of the few that had a true experimental design, in which important alternative factors that might have accounted for positive results in other studies -- different teachers for students with and without computers and different amounts of instructional time students had in the subject area on which they were tested -- were held constant.

The New York City Evaluation of Multiple Integrated Systems

The final entry in Table 1 is a summary of a 400-page report of the effectiveness of 13 different I.L.S. systems placed in service in 26 schools in New York City for the 1987-88 school year (New York City Public Schools, 1989). This potentially powerful evaluation implementation was, however, made impotent by a report that made results presented nearly completely uninterpretable. First, the report failed to identify the nature of the scores shown. Were they raw scores on the same or different levels of the criterion test; scale scores--which show improvement from year to year anyway;--or percentiles or N.C.E. scores--which, as we have shown, have some "natural" gain but not as much as scale scores or raw scores on the same test? Second, the report used a completely irrelevant measure of "effect size," one calculated by dividing the tested cohort's gain from Spring to Spring (on whatever metric was used) by the standard deviation of the gain that individual students made from year to year. That denominator has nothing to do with the standard deviation of individual scores on the statistic (raw score, N.C.E., etc.) itself.

The math scores as shown had numbers greater than 100, indicating that the scores were probably "scale" scores, representing achievement on an absolute scale, and on which students will naturally score higher than they did one year earlier. Whether they scored higher than one would expect them to gain during the year cannot be shown without a comparison or norm population. The reading scores shown, whose school averages typically ranged from 20 to 60, could be percentiles or raw scores or N.C.E.'s. Assuming they were N.C.E.'s, the results show an overall gain of about .2 s.d. Estimating for specific I.L.S.'s is problematic because most had results for fewer than 50 students, and for those that had data on more students, the number of students on which the data was shown was substantially less than the number of students in the program at that school, raising major questions about data quality.

There were other results in the report that suggested something was terribly wrong about the data analysis--for example, only one I.L.S. system showed any relationship between

achievement gains and the amount of time on the system--and that one showed a positive correlation in one subject and a negative one in the other. But the analysts went on to present dozens of conclusions about each of the systems and each of the schools based on their faulty "effect size" statistics that are essentially not worth the paper they are printed on.

Implications

What implications can we draw from all of this. Clearly, in spite of the record of what seems like dozens of "studies" supporting the use of integrated learning systems, to reach a valid conclusion, it is necessary either to put aside the studies with the greatest methodological weaknesses and/or to try to "correct" or allow for those weaknesses to make some estimates as best we can from the studies that were done. Table 3 presents a summary of the findings of the above analysis.

In Table 3, we see that studies of the Wicat system are highly variable, some showing negligible effects, others showing modest effects, and others showing substantial effects. A typical Wicat study has an effect size of +.17, indicating a modest but positive effect on student achievement. The effectiveness of C.C.C. studies seems remarkably related to how they were publicized. C.C.C. studies contained in reports from the vendor all show substantially positive effect sizes. Those obtained from independent sources, once reanalyzed show modest or negligible effects, more on the order of those found for Wicat studies. Only four studies of the Jostens Basic Learning System were found, but two of those, including the experimental design study done by this author, found substantially positive effects.

Table 3 about here

One conclusion stemming from the analysis done to produce this summary is methodological: simpler analyses, which take only reported effect sizes as given and which incorporate studies of widely varying methodological quality without reanalysis, would be likely to produce very different conclusions, conclusions far more positive about the effectiveness of studied programs, as those given here.

Substantively, the widely varying effect sizes and the modestly positive effect sizes that are typical suggest that (1) results will differ as much based on the different conditions of the study as on the different software packages in use (and on the different methodologies used to design and conduct the analysis); (2) that students generally do at least somewhat better using I.L.S.'s than they would be expected to do; (3) that sometimes the results are substantially superior to what one would expect (but it would be difficult at this time to predict the conditions under which those substantial advantages occur); and (4) that evaluation data--particularly evaluation data as produced and disseminated on studies accomplished to date--are too weak a reed on which school districts should base their acquisition and investment decisions.

It would be especially unfair to use existing data to make generalizations about the effectiveness of competing products. First, the sites and research designs that have been used

to study the effects of one vendor's product do not necessarily reflect the same conditions under which a different vendor's products were studied. Second, the relationship between the curriculum provided by a vendor and the criterion measure used to judge effectiveness may not be the same from vendor to vendor. For example, it is this author's impression that the content of the Jostens Basic Learning System is more removed from the content tested on standardized tests than is the content of the C.C.C. or Wicat basic skills systems.

Moreover, it is most important that each of these programs not be seen either as uniform products or as static products. They are not uniform products because they are implemented in very different ways and under very different circumstances from place to place. Among the sites selecting the identical I.L.S. system, the people placed in charge of the students during the computer lab period differ, the relationship between classroom activity and computer lab activity differs, the use of computer-based reports by classroom teachers differs, the computer time used by students and how that time is structured throughout the week or year differs, and so on. Even more important, none of the products remain the same for long. Implementations of C.C.C. that occurred five to ten years ago will not be the same as implementations that occur this year or next year. The curriculum content coverage changes, the hardware platforms on which the students use programs are changing, and certain aspects of the management and diagnoses systems are changing. To the extent that the effectiveness of the instructional program changes with any of these changes, prior evaluation results become only loosely associated with the results in future implementations.

In my efforts to uncover evaluations that have been done and to re-analyze that data to draw out more justifiable conclusions about the effectiveness of prior implementations, I have observed one important thing: The major I.L.S. vendors are very much interested in learning about the effectiveness of their products. It is actually the vendors more than the school districts officials who are trying hardest to accomplish high quality research designs, and who want the studies done to reflect accurate conclusions. The effort to improve the quality of evidence about the effectiveness of these programs is largely a battle between researchers/evaluators in and out of school systems and school system administrators in line authority positions (e.g., superintendents to principals) who respond to a broader set of political and practical considerations. I believe that, given the high financial stakes involved, over time it will become clearer even to those officials that the effort required for more valid research designs and analyses should be given greater support. At least I hope so.

Figure 1. Percentile Gains in National Field Experiment of Computer Use in Math.

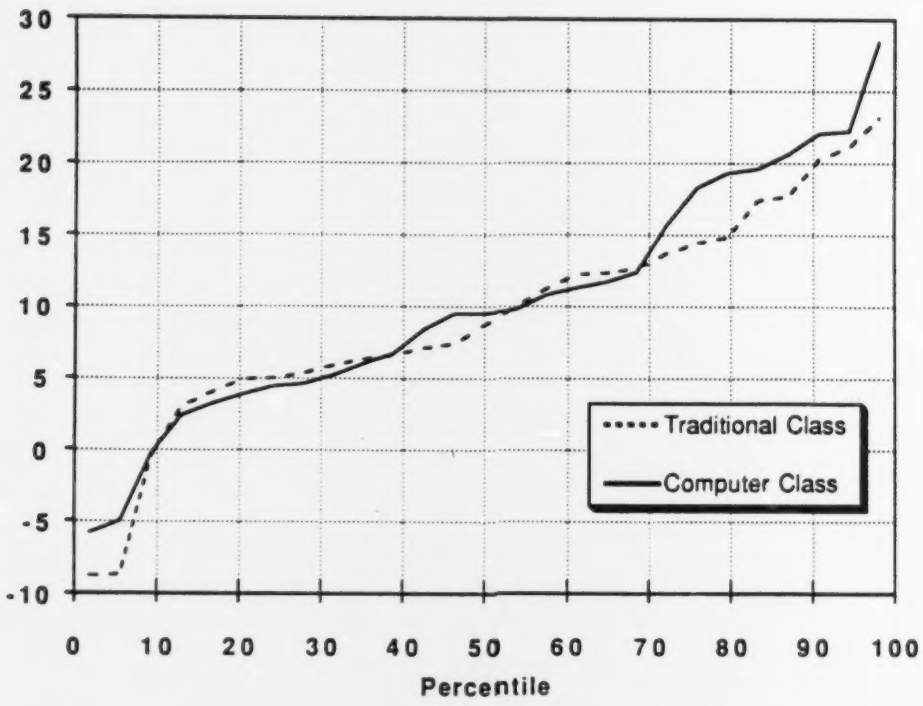

Table 1: Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Chicago, Beale, 1987-88 Wicat School thru Vendor Previous Cohorts' GE Gains	50 students grade 5-6 Chapt. I, "2 lowest-performing classes," low-income black Est. 100 class hours	Spring-Spring ITBS gain: 1.15 grade-equivalents (GE). vs. "previous years" 0.75 GE gain. "Relative improvement" over previous cohorts' gains: 50%	Depends on estimate of s.d. of GE. Using 0.75 to 1.25 years: +.3 < E.S. < +.5	Unclear number or choice of previous years included. No data on cohort comparability (were previous years also for "2 lowest performing classes"?) for this small population. (Same problem for next 3 studies.)	No change but unclear whether comparable to comparison group: E.S. ~ +.4
Chicago, DuBois, 1987-88 Wicat School thru Vendor Previous Cohorts' GE Gains	56 students grades 1-8 Chapt I, low-income minority Est. 35 class hours	December-April ITBS gain; 1.0 years. Relative improvement over previous cohorts' gains which were about .7 years: 46%.	Assuming s.d. of GE is 12 mo. E.S. ~ +.3 in just 5 months (compared to full year previously ???)	Only 56 students tested? Why these? School was large.	No change, but questionable sample of tested students? E.S. ~ +.3
Chicago, Colman, 1985-88 (3 years use) Wicat School thru Vendor Previous Cohorts' GE Gains	500 students grade K-8 (all tested?) Chapt. I, low-income black Est. 75 class hours	Spring-Spring ITBS gain: 10.2 mo. (math); 8.5 mo. (reading). Similar relative improvement over previous cohorts' gains: ~ 60%.	Assuming s.d. of GE is 12 mo. E.S. ~ +.4	Comparison years' GE gain unreasonably low (5.5 mo, math; 4.2 mo, reading). (Typical in inner-city schools avg'g 25th percentile would be about 9 months, same as comparison for study above.)	Using 9 mo. gain as comparison point: E.S. ~ +.1 (math) E.S. ~ .0 (reading)
Chicago, McCorkle, 1986-87 Wicat School thru Vendor Previous Cohorts' GE Gains	400 students grades 4-8 Chapt. I, low-income black Est. 60 class hours	Spring-Spring ITBS gain; 10.6 mo. (math); 8.1 mo. (reading). Relative improvement over previous cohorts' gains: 152%, 42% respectively.	Assuming s.d. of GE is 12 mo. E.S. ~ +.5 (math) E.S. ~ +.2 (reading)	Same problem with low comparison year gain (4.2 mo, math; 5.7 mo., reading). Need reason why these schools should be so low.	Using 9 mo. gain as comparison point: E.S. ~ +.1 (math) E.S. ~ —.1 (reading)

BEST COPY AVAILABLE

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Baytown TX, San Jacinto, 1986-88 (1.5 yrs. use) Wicat LEA thru Vendor Statewide Year-to-Year Changes at Same Grade	200 students grades 3,5 Chapt. I, Hispanic Est. 30 hrs./60 hrs. (year 1 / 2)	State competency test given in February. Change in percent passing test compared to change in statewide percent passing using Feb., 1986 base year and 1988 "post" year: 37% relative gain vs. statewide 20%.	Change over 2 years: E.S. = +.22 overall (+.32 reading, +.11 math, +.22 writing)	Most gain was during first year after just 5 months use. Better comparison point would be gains statewide for Chapter I or Chapter I/Hispanic. Cohorts at a school might differ from one another: no gain scores shown. (Same problem for next four studies.)	First year gain: E.S. = +.17 Second year over 1st: E.S. = +.05
Baytown TX, Carver-Jones, 1987-88 Wicat LEA thru Vendor Comparison School Year-to-Year Changes at Same Grade	240 students grades 3,5 Chapter I, limited-English-proficient Est. 60 class hours	Feb.-administered state competency test. Change in percent passing compared to comparison school. Gained 13% vs. comparison school 1% drop. Gains highest with limited-English students.	E.S. = +.25 average across math, reading, writing, for both grades.	Comparison school unreasonably poor record. Best comparison would be with statewide Chapt I or Chapt. I/Hispanic, but using same statewide comparison as above:	Compared to statewide change of 20% improvement: E.S. = +.08 overall = +.32 math +.13 reading —.21 writing
Irving TX, Schultze, 1986-88 (1.5 yrs. use) Wicat School thru Vendor District Year-to-Year Changes at Same Grade	200 students grades 3,5 Suburban middle-class Est. total 90 hrs over both years	Feb.-administered state competency test. Change in percent passing compared to district average. Over 2 years: relative gain was 17% vs. 6% for district.	E.S. +.12 average across math, reading, writing, for both grades	District gained much less than state over 2 years (5 percentage points vs. 13, or 6% vs. 20% relative gain).	Compared to statewide change: E.S. = —.06
Garland TX, DISTRICT-WIDE, 1986-88 Wicat LEA thru Vendor Statewide Year-to-Year Changes at Same Grade	5000 students grades 3,5 Suburban, middle-class, 19% minority Est. 75 class hours/yr.	Feb.-administered state competency test. Change in percent passing compared to statewide: 15% relative gain vs. 11% statewide for math,reading.	E.S. = +.08	Writing results omitted from summary.	When writing included: E.S. = —.04

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Pharr-San Juan-Alamo TX, DISTRICT-WIDE, 1986-88 (1.5 yrs. use) Wicat LEA thru Vendor Statewide Year-to-Year Changes at Same Grade	2600 students grades 3,5 90% Hispanic Est. 25 hrs. use before 1st testing; 50 hrs use between 1st and 2nd.	Feb.-administered state competency test. Change in percent passing compared to statewide: 31% relative gain vs. 20% statewide over 2 years.	E.S. = +.42 +.42 math +.55 reading +.30 writing	All gain was in first 5 months of use. Raises questions about district's base year testing conditions.	First year gain: E.S. = .47 Second year gain over 1st: E.S. = —.05
Cleveland Cty NC, Washington Elem 1987-88 Wicat School thru Vendor Fall-Spring Change vs. Test Normed Population	350 students grades 1-6 Chapt. I, rural low income, racially mixed Est. 30 class hours	Fall-Spring CAT improvement about 4.5 NCE points from 51 to 55+.	Depends on s.d. of NCE for this population. If in range of 15 to 20 points: +.2 < E.S. < +.3	Fall-Spring norming is poor. Schools often obtain 8 to 10 percentile point gains (5 N.C.E. in this range).	Using 4 to 5 NCE gain as comparison point: E.S. ~ .0
Stillwater OK, 5 schools, 1986-88 (2 years use) Wicat LEA thru Vendor Two-Year Same-Cohort Change vs. Test Normed Population	300 students grade 5 middle class Est. 60 class hours	Two-year longitudinal grade 3 to grade 5 MAT improvement: 7 percentile points from 65 to 72. (= NCE gain 4.5 points)	Assuming s.d. of NCE = 15 to 18: +.25 < E.S. < +.30	Annually, NCE gains nationwide have been about 1.5 points. Also, this is longitudinal comparison of grade cohorts. Grade retention removes lower-performing replaced by older repeaters.	Both factors, principally the former easily reduce E.S. to E.S. ~ .1
Memphis TN, St. Anne's, 1985-88 Wicat Independent thru Vendor Comparison School GE Gains at Same Grade	150 students grade 2-8 Catholic, middle-income Est. 50 class hours	Fall-Spring ITBS GE gains compared to 2 schools with similar SES backgrounds: 1.2 year gains vs. 0.7 year.	Assuming s.d. of GE is in range of 1.25 - 1.75 year +.3 < E.S. < +.4	Unclear whether comparison schools are also parochial schools.	No change. E.S. ~ +.35

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Port Hueneme CA, Blackstock JH, 1984-85 Wicat LEA thru Vendor Earlier Cohort Percentile (implied comparison to test normed pop.)	110 students grade 6 Heterogeneous but 44% Hispanic Est. 50 class hours	Comparison of single score for two 6th grade cohorts two years apart. 30 percentile point difference (improvement) on Spring ITBS in math, 12 points in reading/language.	Assuming s.d. of NCE is in range of 15 to 18 points and assuming cohorts are identical: +.9 < E.S. < +1.2 (math) E.S. ~ +.4 (reading)	Single score for each cohort rather than change score. Differences between cohorts could be responsible. Information not provided for 1984 6th grade, also a pre-Wicat cohort. Why only 110 students? Given school pop, should be about 300 per year. Spring-Spring national gains 1.5 NCE per year.	Need change data or other support for similarity of two compared cohorts. In addition, missing cohort in between and missing students.
Albany OR, Periwinkle, 1984-85 Wicat School thru Vendor One-Year Same-Cohort Change vs. Test Normed Population	300 students grades 2-5 rural, white, middle-class Est. 20 class hours	Winter-Winter CTBS percentile gain 8 points. Levels not stated.	If near mid-range, about 5 NCE gain. If s.d. is 15 points E.S. ~ +.33	Annual NCE gains of 1.5 points.	Accounting for national gains: E.S. ~ +.23
Dearborn MI, Salina, 1985-88 Wicat School thru Vendor District Year-to-Year Changes at Same Grade	80 students grade 4 (3rd year of Wicat) moderate-low income, 95% Arabic-speaking Est. 60 class hours	Percent passing statewide assessment compared to 3 pre-Wicat cohorts. 34% improvement while district improved only 2%.	E.S. = +.4	Always possible that the school (most disadvantaged in the district and with special needs population) received other special services during past 3 years. Also, cohort comparability across 6 years in doubt (3 pre-, 3 post)	Very questionnably due to computer program alone, but not enough information. (E.S. < +.4)
Sandy UT, Waterford, 1982-87 (5 years) Wicat Vendor personal communication One-Year Same-Cohort Change vs. Test Normed Population	80 students grade 4 (4 cohorts) private middle class	Spring-Spring ITBS increases average 5 percentile points per year from 3rd grade to 4th grade classes (+ higher for 1 outlier). (85th - 90th %tile)	Equals 5 NCE point gains for this population. Assuming s.d. of 15 NCE: E.S. ~ +.33	Unknown changes in cohort composition of private school controlling own admissions.	Accounting for national NCE 1.5 point annual gain: E.S. ~ +.23

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Fort Worth Parochial Schools, 1988-89 CCC Vendor Individual Change vs. Test Normed Population	Est.120 math,read,both grades 1-7 Chapter I, parochial schools Approx. 25 clock hours	Spring-Spring Stanfords, Mean NCE gain 15 points, math; 12 points, reading.	E.S. = +1.6 (math) E.S. = +.7 (reading)	Test scores reported for only 60% of students using program. Much larger effects shown for grade 1 than older grades--and what was pretest for that grade? Excluded 3% negative outliers, but no positive outliers--and all of grade 1 reading was outlier (mean NCE pretest score was compared to over 35 for other grades). Mean national increase 2 NCE for Chapt. I not included.	Incorporating mean national NCE gain and excluding grade 1: E.S. ~ +.8 (math) E.S. ~ +.1 (reading)
Omaha Westside Community Schools, 1988-89 CCC Vendor Individual Change vs. Test Normed Population	Est.170 math,read,both grades 2 - 6 Chapt I in 8 schools Approx. 20 clock hours	Spring-Spring CTBS and SRA (varied by school). Mean NCE gain 14 points, math; 11 points, reading.	E.S. = +1.3	Excluded 3% negative outliers. Huge gain for small number of hours. Questionnable.	Incorporate national mean NCE gain: E.S. = +1.1
Milwaukee Public Schools, 1987-88 CCC Vendor Individual Change vs. Test Normed Population	Est.600 math,read,both grades 2-9 Chapt I in 11 schools Approx. 40 clock hours	Fall-Spring in 3 schools; Spring-Spring in 8 schools; ITBS. Mean NCE gain 9 points.	E.S. = +.8 Spring-Spring schools only: E.S. = +.6	Excluded 8% negative outliers in math and 3% in reading. No positive outliers excluded. Also, high attrition: test data on only 60% of students using CCC program.	With 5% students added in with a mean NCE gain of -20 and incorporating national mean gain for Chapt. I of 2 NCE: E.S. ~ +.35 for Spring-Spring schools
Aiken County SC 1987-88 CCC Vendor Individual Change vs. Test Normed Population	Est.600 math,read,both grades 2-8 Chapt I in 8 schools Approx. 30 clock hours	Different tests, different grades (state test and CTBS), probably Spring-Spring, but unstated. Mean NCE gain 7.5 points.	E.S. = +.7	(Limiting to only grades with CTBS pre- and posttests makes no difference)	Incorporating national NCE Spring-Spring gain for Chapt. I of 2 points: E.S. ~ +.5

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Calvert County MD, 1983-88 CCC LEA and state data Changes from Previous Cohorts' Scores Compared to Changes Statewide	Approx. 1500 students per year; Grades 3,5,8 6 elem., 3 middle schls. rural to suburban middle-class heterogeneous district. Est. average 35-40 class hours per year.	District-reported Fall-Fall CAT scores for 4 years prior to CCC (i.e., interval of 3 years) and 3 years after (4.5 and 1.5 for grade 8). GE change in district totals between 1980 and 1986 (CCC started 1983 (1/85 for grades 6+)): + 1.6 years, compared to statewide change of + 0.9 years.	Assuming s.d. of GE in district = 1.1 (grade 3), 1.6 (grade 5), and 2.4 (grade 8) (estimated from district graphs), and using grade-specific GE data: Gains between 1980 & 1986. Statewide gains as comparison group: E.S. ~ +.1 (grade 3) E.S. ~ +.25 (grade 5) E.S. ~ +.5 (grade 8) (although intervention was only 1.5 years for grade 8 students and 3 years for others).	District gains relative to state were larger PRIOR to implementation: Years Grades 3 and 5 Dist. State diff. 1980-83 +.65 +.47 +.18 1983-86 +.35 +.25 +.10 1986-88 -.23 +.03 -.26 Also, younger families in district are more educated, yielding, for example 1980: gr. 3: +.4 GE vs. state gr. 8: —.4 GE vs. state. So younger cohorts (with more CCC experience) should score higher.	(1): Compare closer years: Year just before CCC vs. 2nd year (1st full year) after CCC (again using statewide as comparison point): E.S. ~ +.15 (grade 3) E.S. ~ .0 (grade 5) E.S. ~ .0 (grade 8) (2): Look longitudinally at grade cohorts: 3rd-to-5th grade GE improvement for 4 classes with 2 years of CCC vs. 2 classes with 0 years CCC: E.S. ~ +.1. 5th-to-8th grade GE improvement for 2 classes with 2.5 to 3 years CCC vs. 3 classes with 0 to .5 years: E.S. ~ —.2 Differences between successive cohorts suggest need for longitudinal study below.
Calvert County MD, 1983-86 CCC Independent thru LEA (same data source as above) Individual Change vs. Test Normed Population	653 students (including 94 Chapter I students) followed for 5 years: 2 prior to CCC and 3 after. Received CCC in grades 4-6 or 5-7	Averaging 2 years of fall CAT (pre-) and 3 years (post), mean NCE gain 5.3 points; mean for Chapter I subsample: 10.3 points.	Assuming s.d. of NCE was 15 points (10 for Chapter I): E.S. ~ +.35 (all) E.S. ~ +1.0 (Chap I)	Of 2 years pre-CCC, the first had much lower scores all around. District-wide, gain (relative to state gain) was +.6 GE between the two "pre" years compared to +.1 between last pre- and first post- and —.1 between 1st and 2nd post years.	Assume pre- was non-stable and use only 1983 as pre- data. Only data available, percentiles for median student, converted to NCE's: gained 2 percentiles from 63.5 to 65.5 (= 1 NCE). Incorporating statewide 3 point NCE gain for 3rd grade (1983) cohort over next 2 years: E.S. = —.13. For Chapt I: 3 percentile point gain from 44.5 to 48 = 1.8 NCE. Including comparison group gain: E.S.=—.12

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Los Angeles, CA, 1977-1980 (3.5 years) CCC Independent Random Assignment to Treatments (Different I.L.S. Subjects) and Comparison with Prior or Successor Cohort	Est. 750 students in four schools, grades 1-6 (math), 4-6 (reading), 3-6 (language). Ethnically mixed, large minority populations, two schools were Title I. Est. 50 class hours per year, some in math or reading/language only ; others split.	Spring-Spring CTBS and curriculum-specific test for four years. Random assignment to math, reading/language, or mixed, keeping assignment for up to 4 years. One comparison was between users in one subject and users only in other subject, on tests of both subjects. A 2nd comparison was with grade cohorts immediately earlier or later that did not use CCC. (A 3rd comparison was with students in 2 schools that lacked CCC, but its results are not examined here.)	Reported E.S. included: CTBS MATH computation: (effects CCC math-using vs. reading/lang-using) 1 year: +.36; 3 yrs: +.72; (vs. cohort control) 3 yrs: +.39 CTBS total math: 1 year: +.27; 3 yrs: +.58 vs. cohort controls: +.35 CTBS total reading: 1 year: +.25; 3 yrs: +.12 vs. cohort controls: +.25 CTBS total language 1 year: +.19; 3 yrs: +.26 vs. cohort controls: +.28 Curric.-specific tests: 1 yr: math:+.8; rdg.+.4; lang:+.7	Basically well-designed, but: (1) some students with limited English proficiency removed from Reading and Language groups making groups non-equivalent. (2) Heavy level of support for implementation may make results less generalizable. (3) Cohort comparison may be flawed in that opposite-subject users did better than control in grade 3-5 study. (4) Biggest technical problem is that researchers used residual s.d. rather than raw control group posttest s.d. producing upward tilt to E.S., particularly for reading and language. (5) Because total time in subject not balanced between math-using and reading and language-using students, test results biased in favor of computer users for that subject.	E.S.'s adjusted downwards by using raw control group s.d. for comparability to other studies in education using E.S. measure: E.S. math computation: (1 yr): +.26 (3 yrs): +.59 (cohort): +.33 E.S. math total: (1 yr): +.17 (3 yrs): +.27 (cohort): +.27 E.S. reading (1 yr): +.12 (3 yrs): +.06 (cohort): +.12 E.S. language (1 yr):+.11 (3 yrs): +.14 (cohort): +.19
Lafayette Parish LA, 1980-81 CCC LEA Random Assignment to Treatment (ILS vs. control); individual matches.	94 students at 2 chapter I schools, grades 3 to 6. Math only. Est. 25 class hours.	Spring-Spring CTBS. Students randomly assigned with pretest-matched students at same schools. GE gain 1.01 vs. .88 for control group. Statistically significant difference.	E.S. = +.19	(Only other randomized design of CCC besides L.A. study.)	No change: E.S. = +.2 (math only)

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Portland OR, 1981-82 CCC Independent Comparison Students' Gains; similar population in other schools.	80 students (math only) grades 5-8 Chapt. I, district-wide Est. 25 class hours in math	Fall-Spring LEA assessment. Compared against Chapt I in traditional pull-out in other schools. F-ratio computed from raw scores favored CCC. Students tested after summer: difference no longer statistically significant	E.S. = +.30 (fall-spr) E.S. = +.13 (fall-fall)	Traditional Chapt I taught in elementary schools; CCC in middle schools; Traditional averaged about .5 years younger grade level. But groups pretested evenly, indicating CCC group pretested relatively lower for their grade.	No change: E.S. = +.3
Rochester NY, 1984-86 CCC LEA Comparison Students' Gains; pretest matched.	1200 students, year 1; 2600 students, year 2 grades 4-6 at 19 schools. Pull-out, but not limited to Chapter I.	Spring-Spring MAT. Year 1: individual students matched to non-users; year 2, only rough group match (not successful--non-users higher pretests). No signif. difference either year.	Assuming s.d. of NCE was 9 in reading, 13 in math (from rough data): E.S. year 1: ~ —.06 E.S. year 2: math: ~ .0 reading: ~ —.14	District did poorly in year 1, with or without CCC.)	Overall estimate: .0 < E.S. < —.1
Atlanta, GA 1984-85 CCC LEA Comparison Students' Gains, non-users same school	700 students elem. to H.S. in 11 schools Majority (~60%) were Chapter I or other low-achievers. Est. 25 class hours.	Probably Spring-Spring (but not stated) CAT compared to students at same schools not using CCC. Significantly higher gains by CCC at elem. and H.S. Elem. gained 7 NCE reading; 5 math.	Assuming s.d. of NCE was 12: E.S. ~ +.6 (math) E.S. ~ +.4 (reading)	No data at all on pretest equivalence or explanation how users were chosen. Not explicitly stated that gains were Spring-Spring	Not willing to accept E.S. numbers as valid because of inadequate comparison group, but if data were Spring-Spring, probably was effective.

BEST COPY AVAILABLE

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Woodlands Elem, 1986-87 Jostens Vendor Fall-Spring Gains vs. Test Normed Population	200 students grades 2, 4, 6 middle-class but heavy military population Est. 35 class hours	Fall-Spring MAT. Mean percentile gains 9 points equal to 5 NCE at this range.	Assuming s.d. of NCE was 15 points: E.S. ~ +.33	Fall to spring inadequacies. Must assume typical 8 to 10 percentage point gain as comparison. Cohort comparisons. Large turnover in military-family schools. Unknown comparability of drops and adds.	Incorporating 8 to 10 percentage point gain as comparison: E.S. ~ .0
Lexington, KY, 1987-88 Jostens LEA Comparison Students' Gains, Other Schools	about 250 students in 4 schools. grades 2-6 (reading); 4-6 (math) Chapter I students. Est. 15 class hours	Fall-Spring MAT gains compared to Chapter I students at 11 other schools. Jostens schools gained 4.5 NCE more than comparison schools in math, and 3 more in reading, but all positive reading results were grades 2-3.	Assuming s.d. of NCE in this Chapter I population was 10 points: E.S. ~ +.4 (math) E.S. ~ +.5 (rdg, gds 2-3) E.S. ~ .0 (rdg, gds 4-6)	No data provided about comparability of students in comparison schools; e.g., pretest scores (only gains shown). Small number of hours to produce large gains. Questionnable.	Need more data on comparability, but possibly... E.S. ~ .4 (math) E.S. ~ .25 (avg.,reading)
Fairfax VA, 1 school, 1987-88 Jostens Independent Random Assignment to Treatment (Different I.L.S. subjects)	240 students using for either math or reading grades 4-6 school with military population, 37% minority Est. 30 class hours in one subject.	Fall-Spring CTBS gains on one subject (e.g., math) were compared with students using Jostens only with other subject (e.g., reading). Scale score gains favored Jostens user for one grade, reading. No other differences.	Using data on raw gains and given s.d.: E.S. ~ +.05 (math) E.S. ~ +.12 (reading)	Large attrition (40%) from original sample size, probably because of high mobility of this population.	No change: E.S. ~ +.05 (math) E.S. ~ +.1 (reading)

Table 1 (cont.): Studies of Integrated Learning Systems in Elementary and Middle Grades

Location, Year, Program, Source of Data Research Design: Comparison Group	Approx. # of ILS-Using Students Tested; Grade Levels; Population; Hours of Use of ILS (math + language + reading) unless noted	Summary of Effects Either Reported by Evaluator or Found in Evaluator's Report	Effect Size (E.S.) Indicated by that Summary	Problem with Comparison and Other Comments Regarding Study	Alternative Effect Size
Hurst TX, Hurst Hills, 1987-89 Jostens Independent Random Assignment to Treatment (I.L.S. vs. traditional), Both Treatments Taught by Same Teacher	67 students (math only grades 5,6 Suburban, heterogeneous. 57 class hours in one subject.	Combines 2 successive years' classes: Year 1: Fall CAT to Spring Stanfords & curriculum-specific and oral tests. Year 2: same but Fall Stanfords. Pretest controlled raw scores compared to control groups'.	Reported by investigator (this author): E.S. = +.45 (total math) = +.28 (computation) = +.53 (applications) = +.49 (curric.spec.) = +.76 (estimation/ mental math)	(Out of 29 schools with 1 to 4 pairs of randomized classes (about half randomized at student level; rest supposed to be similar), this school had highest mean E.S. and was only 1 of 2 above +.2. Two teachers here: one had highest E.S. both years among 50 pairs for computation; other had two of three highest among 50 pairs on curriculum-specific test.)	No change: E.S. = +.45
New York City, 1987-88 Many programs LEA Individual Gains vs. ???	from under 50 to 600 students per ILS, each in one to four schools; mixed grade levels About 50% Chapter I, others in other programs and general urban population Est. 10 hours clock time	Spring-Spring Degrees of Reading Power and MAT Math. Mean relative gain in scores (not clear if scores are NCE, percentile, scale scores or what) was roughly equal to .9 times the s.d. of the gain. (That indicates only the shape of the gain curve--flat or steep.)	"Effect size," as used by the investigators, does not mean the same thing as the term is commonly used. The value "E.S. = +.9" in the report has no relation to an E.S. in the other studies.	If the reading scores, which average about "48" (1987) and "53" (1988) can be assumed to be NCE's, then could estimate. Assume s.d. of NCE is 14 for this population. Could estimate for some ILS's with sufficient N. CCC is one: but numbers of cases shown is only 35% of original number shown for CCC. No data on standard deviations of math scores. Cannot calculate E.S. Little clock time to expect results anyway. Different conditions in different schools--as with most comparison school designs.	Assuming reading scores are NCE's, and incorporating 2 point NCE national annual gain, average over all I.L.S.'s: E.S. ~ +.2 (If numbers were percentiles, E.S. ~ +.1)

BEST COPY AVAILABLE

Table 2: Calvert County, MD., and Maryland State-wide Changes in Mean Grade Equivalent Scores, 1980-88, California Achievement Test*

	3rd Grade			5th Grade			8th Grade		
	District	State	(diff)	District	State	(diff)	District	State	(diff)
Gains prior to IIS									
1980-83	+.37	+.43	-.06	+93	+.50	+.43	+1.67	+1.03	+.64
Gains after IIS									
1983-86	+.33	+.13	+.20	+.37	+.37	.00			
1986-88	-.20	+.03	-.23	-.20	.00	-.20			
1983-88	+.13	+.16	-.03	+.17	+.37	-.20			
1985-88 **							+.13	+.30	-.17

* All tests are fall of each year.

** Implemented in middle of 1984-85 school year, so fall 1985 is not exactly a baseline point. If fall 1984 is used instead, comparable gains are: district: +.30; state: +.43

Table 3: Summary of Effect Size Analysis of Integrated Learning Systems

Program and Source	E.S. Under +.15 (Negligible)	E.S. Betw. +.15 and +.30 (Moderate)	E.S. Over +.30 (Substantial)	Median Effect Size	E.S. of Randomized Designs
Wicat through Vendor	✓✓✓✓ ✓✓✓	✓✓✓✓	✓✓✓✓	+.17	-
CCC through Vendor	-	-	✓✓✓✓	+.40	-
CCC: LEA or Independent	✓✓	✓✓✓	-	+.15	+.17 (2)
Jostens through Vendor	✓	-	-	+.0	-
Jostens Independent	✓	-	✓✓	+.33	+.26 (2)

References

Alifrangis, C. M. (1989), A Critical Analysis of the Components of an Integrated Learning System and a Measure of the System's Effect on Mathematics and Reading Achievement of Fourth through Sixth Grade Students (Doctoral dissertation, George Mason University).

Anderson, J. I. (1984), Technical issues in measuring program impact: The reason for alternatives, Paper presented at the Annual Meeting of the American Educational Research Association.

Atlanta Public Schools (1984), The Evaluation of the Atlanta Public Schools Computer Curriculum Corporation Labs: 1984-85, Atlanta, Georgia.

Austin, G. R. (1988), Computer-Assisted Instruction, Calvert County, Maryland Public Schools 1983-1987, University of Maryland Baltimore County Center for Educational Research and Development.

Austin, G. R., & T. Howie (1990), Instructional Leadership, Computer-Aided Instruction, and Sustained Effects, University of Maryland Baltimore County.

Becker, H. J. (1986a), Instructional Uses of School Computers: Reports from the 1985 national survey, Johns Hopkins University Center for Social Organization of Schools, Issue No. 3, November.

Becker, H. J. (1986b), The Effects of Computer Use on Children's Learning: Limitations of Past Research and a Working Model for New Research, Peabody Journal of Education, 64:1, pp. 81-110.

Becker, H. J. (1990a), Computer Use in United States Schools, 1989: An Initial Report of U.S. Participation in the I.E.A. Computers in Education Survey, Paper presented at the Annual Meeting of the American Educational Research Association.

Becker, H. J. (1990b), Effects of Computer Use on Mathematics Achievement: Findings from a Nationwide Field Experiment in Grade Five to Eight Classes, Paper presented at the Annual Meeting of the American Educational Research Association.

Calvert County Public Schools (no date), Choosing for the Future, Calvert County, Maryland.

Cannell, J. J. (1988), How Public Educators Cheat on Standardized Achievement Tests, Friends for Education, pp. 43-49, 107-109.

EPIE Instituite (1990), The Integrated Instructional Systems Report, Water Mill, New York.

Gabriel, R. M., B. L. Anderson, G. Benson, S. Gordon, R. K. Hill, J. Pfannenstiel, R. M. Stonehill (1985), Studying the Sustained Achievement of Chapter 1 Students, U. S. Depatment of Education, Washington, D.C.

Gevirtz, G., & P. Kelman (1990), The Scholastic Guide to Educational Computer Networks, New York, New York.

Gourgey, A, J. Azumi, S. Madhere, & E. Walker (1984), Computer-assisted instruction evaluation report, 1983-1984, Newark School District, Newark, New Jersey.

Hotard, S. R., & M. J. Cortez (no date), Computer Assisted Instruction as an Enhancer of Remediation, University of Southwestern Louisiana.

Kitchen, D. J. (1986), Computer-Assisted Mathematics Remediation: A First-Year Evaluation of the Computer Curriculum Corporation System in a Senior High School Setting, San Diego State University.

Maryland State Depatment of Education (1989), Maryland Accountability Testing Program Annual Report: School Year 1988-1989, Baltimore, Maryland.

Miller, S. (1984), A comparison of computer-assisted instruction with prescription learning and the traditional "pull out" program used in supplementing instruction of mathematics basic skills to chapter 1 students (Doctoral dissertation, Oregon State University).

New York City Public Schools (1989), Computer Pilot Program: 1987-1988, New York, New York.

Ragosta, M. (1983), Computer-Assisted Instruction and Compensatory Education: A Longitudinal Analysis, Machine-Mediated Learning, 1:1, pp. 97-127.

Ragosta, M., P. W. Holland, & D. T. Jamison (1982), Computer-Assisted Instruction and Compensatory Education: The ETS/LAUSD Study, Educational Testing Service, Princeton, New Jersey.

Sacramento Unified School District (1989), An Evaluation of the Compensatory Education Computer Assisted Instruction Program for Chapter I Students: Kindergarten through Grade 6, Sacramento, California.

Shattuck, J. (1986), Computer-assisted instruction: Evaluation Summary, Rochester City School District, Rochester, New York.

Suppes, P., M. Zanotti, & N. Smith (1988a), Effectiveness of the CCC CAI Program for EIA Remedial and Compensatory Program in Aiken County, South Carolina, Computer Curriculum Corporation, Palo Alto, California.

Suppes, P., M. Zanotti, & N. Smith (1988b), Effectiveness of the CCC CAI Program: Milwaukee Public Schools, Computer Curriculum Corporation, Palo Alto, California.

Suppes, P., M. Zanotti, & N. Smith (1989a), Effectiveness of the CCC CAI Program: Corpus Christi Parochial Schools, Computer Curriculum Corporation, Palo Alto, California.

Suppes, P., M. Zanotti, & N. Smith (1989b), Effectiveness of the CCC CAI Program: Fort Worth Parochial Schools, Computer Curriculum Corporation, Palo Alto, California.

Suppes, P., M. Zanotti, & N. Smith (1989c), Effectiveness of the CCC CAI Program: Omaha Westside Community Schools, Computer Curriculum Corporation, Palo Alto, California.

WICATsystems (no date), Learning Improvement Reports, Orem, Utah.

Zollman, A., B. Oldham, & J. Wyrick (1989), Effects of Computer-Assisted Instruction on Reading and Mathematics Achievement of Chapter 1 Students, Fayette County Public Schools Printing Office, Lexington, Kentucky.